Metrical Tagging in the Wild:
Building and Annotating Poetry Corpora with Rhythmic Features

Thomas Haider

Department of Language and Literature,
Max Planck Institute for Empirical Aesthetics, Frankfurt am Main
Institute for Natural Language Processing (IMS),
University of Stuttgart

`thomas.haider@ae.mpg.de`

Abstract

A prerequisite for the computational study of literature is the availability of properly digitized texts, ideally with reliable meta-data and ground-truth annotation. Poetry corpora do exist for a number of languages, but larger collections lack consistency and are encoded in various standards, while annotated corpora are typically constrained to a particular genre and/or were designed for the analysis of certain linguistic features (like rhyme). In this work, we provide large poetry corpora for English and German, and annotate prosodic features in smaller corpora to train corpus driven neural models that enable robust large scale analysis.

We show that BiLSTM-CRF models with syllable embeddings outperform a CRF baseline and different BERT-based approaches. In a multi-task setup, particular beneficial task relations illustrate the inter-dependence of poetic features. A model learns foot boundaries better when jointly predicting syllable stress, aesthetic emotions and verse measures benefit from each other, and we find that caesuras are quite dependent on syntax and also integral to shaping the overall measure of the line.

1 Introduction

Metrical verse, lyric as well as epic, was already common in preliterate cultures (Beissinger, 2012), and to this day the majority of poetry across the world is drafted in verse (Fabb and Halle, 2008). In order to reconstruct such oral traditions, literary scholars mainly study textual resources (rather than audio). The rhythmical analysis of poetic verse is still widely carried out by example- and theory-driven manual annotation of experts, through so-called close reading (Carper and Attridge, 2020; Kiparsky, 2020; Attridge, 2014; Menninghaus et al., 2017). Fortunately, well-defined constraints and the regularity of metrically bound language aid the prosodic interpretation of poetry.

However, for projects that work with larger text corpora, close reading and extensive manual annotation are neither practical nor affordable. While the speech processing community explores end-to-end methods to detect and control the overall personal and emotional aspects of speech, including fine-grained features like pitch, tone, speech rate, cadence, and accent (Valle et al., 2020), applied linguists and digital humanists still rely on rule-based tools (Plecháč, 2020; Anttila and Heuser, 2016; Kraxenberger and Menninghaus, 2016), some with limited generality (Navarro-Colorado, 2018; Navarro et al., 2016), or without proper evaluation (Bobenhausen, 2011). Other approaches to computational prosody make use of lexical resources with stress annotation, such as the CMU dictionary (Hopkins and Kiela, 2017; Ghazvininejad et al., 2016), are based on words in prose rather than syllables in poetry (Talman et al., 2019; Nenkova et al., 2007), are in need of an aligned audio signal (Rosenberg, 2010; Rösiger and Riester, 2015), or only model narrow domains such as iambic pentameter (Greene et al., 2010; Hopkins and Kiela, 2017; Lau et al., 2018) or Middle High German (Estes and Hench, 2016).

To overcome the limitations of these approaches, we propose corpus driven neural models that model the prosodic features of syllables, and to evaluate against rhythmically diverse data, not only on syllable level, but also on line level. Additionally, even though practically every culture has a rich heritage of poetic writing, large comprehensive collections of poetry are rare. We present in this work datasets of annotated verse for a varied sample of around 7000 lines for German and English. Moreover, we collect and automatically annotate large poetry corpora for both languages to advance computational work on literature and rhythm. This may include the analysis and generation of poetry, but also more general work on prosody, or even speech synthesis.

Our main contributions are:

1. The collection and standardization of heterogenous text sources that span writing of the last 400 years for both English and German, together comprising over 5 million lines of poetry.

2. The annotation of prosodic features in a diverse sample of smaller corpora, including metrical and rhythmical features and the development of regular expressions to determine verse measure labels.

3. The development of preprocessing tools and sequence tagging models to jointly learn our annotations in a multi-task setup, highlighting the relationships of poetic features.

2   Manual Annotation

We annotate prosodic features in two small poetry corpora that were previously collected and annotated for aesthetic emotions by Haider et al. (2020). Both corpora cover a time period from around 1600 to 1930 CE, thus encompassing public domain literature from the modern period. The English corpus contains 64 poems with 1212 lines. The German corpus, after removing poems that do not permit a metrical analysis, contains 153 poems with 3489 lines in total. Both corpora are annotated with some metadata such as the title of a poem and the name and dates of birth and death of its author. The German corpus further contains annotation on the year of publication and literary periods.

Figure 1 illustrates our annotation layers with three fairly common ways in which poetic lines can be arranged in modern English. A poetic line is also typically called *verse*, from Lat. *versus*, originally meaning to turn a plow at the ends of successive furrows, which, by analogy, suggests lines of writing (Steele, 2012).

In this work, we manually annotate the sequence of syllables for metrical (meter, met) prominence (+/−), including a grouping of recurring metrical patterns, i.e., foot boundaries (|). We also operationalize a more natural speech rhythm (rhy) by annotating pauses in speech, caesuras (:), that segment the verse into rhythmic groups, and in these groups we assign main accents (2), side accents (1) and null accents (0). In addition, we develop a set of regular expressions that derive the verse measure (msr) of a line from its raw metrical annotation.

In English or German, the rhythm of a linguistic utterance is basically determined by the sequence of syllable-related accent values (associated with pitch, duration and volume/loudness values) resulting from the 'natural' pronunciation of a line, sentence or text by a competent speaker who takes into account the learned inherent word accents as well as syntax- and discourse-driven accents. Thus, lexical material comes with n-ary degrees of stress, depending on morphological, syntactic, and information structural context. The prominence (or stress) of a syllable is thereby dependent on other syllables in its vicinity, such that a syllable is pronounced relatively louder, higher pitched, or longer than its adjacent syllable.

```
msr iambic.pentameter
met −   + | −    + | −    + | −  +| − +  |
rhy 0   1   0   0   0   2 : 0  1  0  2  :
    My love is like to ice, and I to fire:

msr iambic.tetrameter
met −   + |−   + | −   + | −   +  |
rhy 0   1  0   2  0   0  0   1  :
    The winter evening settles down

msr trochaic.tetrameter
met +    −  | +   − | + − | +
rhy 0    0  1   0  1  0   2  :
    Walk the deck my Captain lies,
```

Figure 1: Examples of rhythmically annotated poetic lines, with meter (+/−), feet (|), main accents (2,1,0), caesuras (:), and verse measures (msr). Authors: Edmund Spenser, T.S. Eliot, and Walt Whitman.

2.1   Annotation Workflow

Prosodic annotation allows for a certain amount of freedom of interpretation and (contextual) ambiguity, where several interpretations can be equally plausible. The eventual quality of annotated data can rest on a multitude of factors, such as the extent of training of annotators, the annotation environment, the choice of categories to annotate, and the personal preference of subjects (Mo et al., 2008; Kakouros et al., 2016).

Three university students of linguistics/literature were involved in the manual annotation process. They annotated by silent reading of the poetry, largely following an intuitive notion of speech rhythm, as was the mode of operation in related work (Estes and Hench, 2016). The annotators additionally incorporated philological knowledge to recognize instances of poetic license, i.e., knowing how the piece is supposed to be read. Especially

the annotation accuracy of metrical syllable stress and foot boundaries benefited from recognizing the schematic consistency of repeated verse measures, license through rhyme, or particular stanza forms.

2.2 Annotation Layers

In this paper, we incorporate both a linguistic-systematic and a historically-intentional analysis (Mellmann, 2007), aiming at a systematic linguistic description of the prosodic features of poetic texts, but also using labels that are borrowed from historically grown traditions to describe certain forms or patterns (such as verse measure labels).

We evaluated our annotation by calculating Cohen's Kappa between annotators. To capture different granularities of correctness, we calculated agreement on syllable level (accent/stress), between syllables (for foot or caesura), and on full lines (whether the entire line sequence is correct given a certain feature).

Main Accents & Caesuras: Caesuras are pauses in speech. While a caesura at the end of a line is the norm (to pause at the line break) there are often natural pauses in the middle of a line. In few cases the line might also run on without a pause. As can be seen in Figure 1, punctuation is a good signal for caesuras. Caesuras (csr) are denoted with a colon. We operationalize rhythm by annotating three degrees of syllable stress, where the verse is first segmented into rhythmic groups by annotating caesuras, and in these groups we assign main accents (2), side accents (1) and null accents (0).

	Syllable		Whole Line	
	m.ac	caesura	m.ac	caesura
DE_{blind}	.84	.92	.59	.89
EN_{blind}	.80	.88	.66	.86

Table 1: Cohen Kappa Agreement for Main Accents and Caesura

Six German and ten English poems were annotated by two annotators to calculate the agreement for rhythm. Table 1 lists the agreement figures for main accents (m.ac) and caesuras. It shows that caesuras can be fairly reliably detected through silent reading in both languages. On the other hand, agreement on main accents is challenging. Figure 2 shows the confusion of main accents for German.

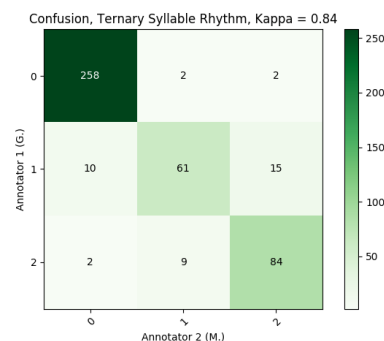While 0s are quite unambiguous, it is not always clear when to set a primary (2) or side accent (1).

Figure 2: Confusion of German Main Accents

Meter and Foot: In poetry, meter is the basic prosodic structure of a verse. The underlying abstract, and often top-down prescribed, meter consists of a sequence of beat-bearing units (syllables) that are either prominent or non-prominent. Non-prominent beats are attached to prominent ones to build metrical feet (e.g. iambic or trochaic ones). This metrical structure is the scaffold, as it were, for the linguistic rhythm. Annotators first annotated the stress of syllables and in a subsequent step determined groupings of these syllables with foot boundaries, thus a foot is the grouping of metrical syllables. The meter (or measure) of a verse can be described as a regular sequence of feet, according to a specific sequence of syllable stress values.

	Syllable		Whole Line	
	meter	foot	meter	foot
$DE_{corr.}$	.98	.87	.94	.71
DE_{blind}	.98	.79	.92	.71
EN_{blind}	.94	.95	.87	.88

Table 2: Cohen Kappa Agreement for Metrical Stress and Foot Boundaries. Corr. is the agreement of the first version against the corrected version. Blind means that annotators did not see another annotation.

The meter annotation for the German data was first done in a full pass by a graduate student. A second student then started correcting this annotation with frequent discussions with the first author. While on average the agreement scores for all levels of annotation suggested reliable annotation after an initial batch of of 20 German poems, we found that agreement on particular poems was far lower

3717

than the average, especially for foot boundaries. Therefore, we corrected the whole set of 153 German poems, and the first author did a final pass. The agreement of this corrected version with the first version is shown in Table 2 in the row $\text{DE}_{corr.}$. To check whether annotators also agree when not exposed to pre-annotated data, a third annotator and the second annotator each annotated 10 diverse German poems from scratch. This is shown in DE_{blind}. For English, annotators 2 and 3 annotated 6 poems blind and then split the corpus.

Notably, agreement on syllables is acceptable, but feet were a bit problematic, especially for German. To investigate the sources of disagreement, we annotated and calculated agreement on all 153 poems. Close reading for disagreement of foot boundaries revealed that poems with κ around .8 had faulty guideline application (annotation error). 14 poems had an overall $\kappa < .6$, which stemmed from ambiguous rhythmical structure (multiple annotations are acceptable) and/or schema invariance, where a philological eye considers the whole structure of the poem and a naive annotation approach does not render the intended prosody correctly.

As an example for ambiguous foot boundaries, the following poem, Schiller's 'Bürgschaft', can be set in either *amphibrachic* feet, or as a mixture of *iambic* and *anapaestic* feet. Such conflicting annotations were discussed by Heyse (1827), who finds that in the Greek tradition the *anapaest* is preferable, but a 'weak amphibrachic gait' allows for a freer rhythmic composition. This suggests that Schiller was breaking with tradition.

```
(Foot Boundary Ambiguity) Schiller, 'Die Bürgschaft'

(1) met="-+-|-+-|-+-|"
    Ich lasse |   den Freund   dir | als Bürgen, |
(2) met="-+|--+|--+|-"
    Ich las   | se den Freund | dir  als Bürg | en,
Transl.: I leave this friend to you as guarantor

(1) met="-+-|-+-|-+-|"
    Ihn magst   du, | entrinn'   ich, | erwürgen. |
(2) met="-+|--+|--+|-"
    Ihn magst | du,   entrinn' | ich,   erwür | gen.
Transl.: Him you may strangle if I escape.

(1) (amphibrach)
(2) (iambus / anapaest)
```

Verse Measures: We develop a set of regular expressions to determine the measure of a line from its raw metrical annotation. We orient ourselves with the handbook of Knörrich (1971). The 'verse measure' (msr) is a label for the whole line according to recurring metrical feet. We label the verse according to its dominant foot, i.e., the repetition of patterns like iambus (-+),

trochee (+-), dactyl (+--), anapaest (--+), or amphibrach (-+-). Also, the rules determine the number of stressed syllables in the line, where *di-, tri-, tetra-, penta-,* and *hexameter* signify 2, 3, 4, 5, and 6 stressed syllables accordingly. Thus, +-+-+- is an example for a trochaic.trimeter and -+-+-+-+ is a iambic.tetrameter, since the foot boundaries should look like this: -+|-+|-+|-+|. Typically, female (unstressed) line endings are optional (cadence). Additionally, we annotate labels for (i) inversion, when the first foot is inverted, e.g., the first foot in a iambic line is trochaic: +--+-+-+, (ii) relaxed, if an unstressed syllable was inserted: -+-+--+-+ (iambic.tetrameter.relaxed), (iii) and choliambic endings: -+-+-+--+. Besides these basic forms, we also implement historically important forms such as a strict alexandrine,[1] the dactylic hexameter,[2] conventionally known as 'hexameter', and some ode forms like the asklepiadic verse (+-+--++--+-+).

Table 3 lists the most frequent labels for each language without length, called short measure (smsr). The English data includes all datasets that are used in the experiments, as discussed in section 4.1.3.

English		||	German	
freq.	smsr	||	freq.	smsr	
2096	iambic	||	1976	iambic	
490	trochaic	||	793	trochaic	
306	anapaest	||	258	amphibrach	
255	amphibrach	||	206	alexandrine	
248	daktylic	||	76	daktylic	
152	hexameter	||	72	anapaest	
91	prosodiakos	||	26	asklepiade	
52	other	||	17	pherekrateus	
35	alexandrine	||	14	glykoneus	

Table 3: Most frequent verse measures in small English and German corpora, without length.

3 Large Poetry Corpora

In order to enable large scale experiments on poetry, we collect and standardize large poetry corpora for English and German. The English corpus contains around 3 million lines, while the German corpus contains around 2 million lines. The corpora and code can be found at https://github.com/tnhaider/metrical-tagging-in-the-wild

---

[1] Alexandrine: -+-+-+-+-+-? The symbol before ? is optional

[2] Hexameter: +--?+--?+--?+--?+--+-

3718

Our resources are designed in a standardized format to sustainably and interoperably archive poetry in both .json and TEI P5 XML. The .json format is intended for ease of use and speed of processing while retaining some expressiveness. Our XML format is built on top of a "Base Format", the so-called DTA-Basisformat[3] (Haaf et al., 2014) that not only constrains the data to TEI P5 guidelines, but also regarding a stricter relaxNG schema that we modified for our annotation.[4]
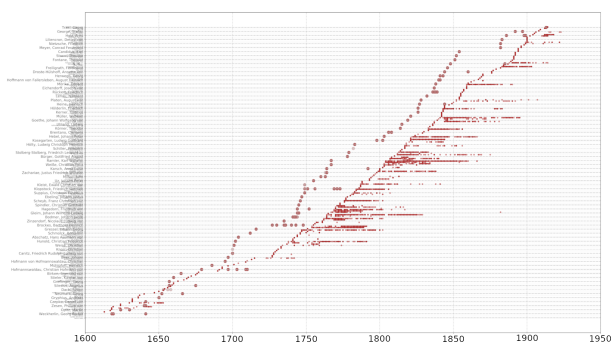
3.1 A Large German Poetry Corpus

Figure 3: Each dot represents a poem of a German Author (y-axis) over Time (x-axis) from Textgrid (small dots bottom) and DTA (large dots top). 1600–1950. Authors are not aligned, and poems can be on top of each other.

We built a large, comprehensive, and easily searchable resource of New High German poetry by collecting and parsing the bulk of digitized corpora that contain public domain German literature. This includes the German Text Archive (DTA) (`http://deutschestextarchiv.de`) the Digital Library of Textgrid (`http://textgrid.de`), and also the German version of Project Gutenberg (which we omit from our experiments due to inconsistency).[5]

Each of these text collections is encoded with different conventions and varying degrees of consistency. Textgrid contains 51,264 poems with the genre label 'Verse', while DTA contains 23,877 poems with the genre label 'Lyrik'. It should be noted that the whole DTA corpus contains in total 40,077 line groups that look like poems, but without the proper genre label, poems are likely embedded within other texts and might not come with proper meta-data. We implement XML parsers in python to extract each poem with its metadata and fix stanza and line boundaries. The metadata includes the author name, the title of the text, the year it was published, the title and genre of the volume it was published in, and finally, an identifier to retrieve the original source. We perform a cleaning procedure that removes extant XML information, obvious OCR mistakes, and normalize umlauts and special characters in various encodings,[6] particularly in DTA. We use langdetect[7] 1.0.8 to tag every poem with its language to filter out any poems that are not German (such as Latin or French). The corpus finally almost 2M lines in over 60k poems. In Figure 3 we plotted each poem in DTA and Textgrid over time, from 1600 to 1950. The x-axis shows the year of a poem, while the y-axis is populated by authors. One can see that DTA consists of full books that are organized by author (large dots) so that the datapoints for single poems get plotted on top of each other, while Textgrid has a time stamp for most single poems (after 1750), outlining the productive periods of authors.

3.2 A Large English Poetry Corpus

The English corpus contains the entirety of poetry that is available in the English Project Gutenberg (EPG) collection. We firstly collected all files with the metadatum 'poetry' in (temporal) batches with the GutenTag tool (Brooke et al., 2015), to then parse the entire collection in order to standardize the inconsistent XML annotation of GutenTag and remove duplicates, since EPG contains numerous different editions and issues containing the same material. We also filter out any lines (or tokens) that indicate illustrations, stage directions and the like. We use langdetect to filter any non-English material.

The github repository of Parrish (2018) previously provided the poetry in EPG by filtering single lines with a simple heuristic (anything that could look like a line), not only including prose with line breaks, but also without conserving the integrity of poems but providing a document identifier per line to find its origin. We offer our corpus in XML with intact document segmentation and metadata, still containing over 2.8 million lines.

---

[3] `http://www.deutschestextarchiv.de/doku/basisformat/`

[4] This schema defines a strict layout of poetic annotation. It allows us to validate XML files regarding their correctness. It is thus useful for manual annotation with the OxygenXML editor, avoiding parsing errors downstream.

[5] `https://www.projekt-gutenberg.org/`

---

[6] We fix the orthography both on string and bytecode level. We replace the rotunda (U+A75B) and the long s (U+017F), the latter of which is pervasive in DTA.

[7] `https://pypi.org/project/langdetect/`

4 Experiments

In the following, we carry out experiments to learn the previously annotated features and determine their degree of informativeness for each other with a multi-task setup. We include two additional datasets with English meter annotation, and evaluate pre-processing models for syllabification and part-of-speech tagging.

4.1 Preprocessing

Tokenization for both languages is performed with SoMaJo with a more conservative handling of apostrophes (to leave words with elided vowels intact) (Proisl and Uhrig, 2016). This tokenizer is more robust regarding special characters than NLTK. We also train models for hyphenation (syllabification) and part-of-speech (POS) tagging, since syllabification is a prerequisite to analyse prosody, and POS annotation allows us to gauge the role of syntax for prosodic analysis.

4.1.1 Hyphenation / Syllabification

For our purposes, proper syllable boundaries are paramount to determine the segmentation of lines regarding their rhythmic units. The test the following systems:[8] *Sonoripy*,[9] *Pyphen*,[10] *hypheNN*,[11] and a BiLSTM-CRF (Reimers and Gurevych, 2017)[12] with pretrained word2vec character embeddings. These embeddings were trained on the corpora in section 3.

To train and test our models, we use CELEX2 for English and extract hyphenation annotation from wiktionary for German.[13] We evaluate our models on 20,000 random held-out words for each language on word accuracy and syllable count. Word accuracy rejects any word with imperfect character boundaries, while syllable count is the more important figure to determine the proper length of a line. As seen in Table 4, the BiLSTM-CRF

---

[8]Syllabipy determines boundaries based on the sonority principle, Pyphen uses the Hunspell dictionaries, and HypheNN is a simple feed forward network that is trained on character windows.

[9]https://github.com/alexestes/SonoriPy
https://github.com/henchc/syllabipy

[10]pyphen.org

[11]github.com/msiemens/HypheNN-de

[12]https://github.com/UKPLab/emnlp2017-bilstm-cnn-crf

[13]For German, wiktionary contains 398.482 hyphenated words, and 130.000 word forms in CELEX. Unfortunately, German CELEX does not have proper umlauts, and models trained on these were not suitable for poetry. For English, wiktionary only contains 5,142 hyphenated words, but 160,000 word forms in CELEX.

performs best for English and does not need any postprocessing. For German, the LSTM model is less useful as it tends to overfit, where over 10% of annotated lines were still rejected even though in-domain evaluation suggests good performance. We therefore use an ensemble with HypheNN, Pyphen and heuristic corrections for German, with only 3% error on the gold data, as seen in Table 5 (the datasets are discussed in section 4.1.3).

| | German | | English | |
	w. acc.	sy. cnt	w. acc.	sy. cnt
SonoriPy	.476	.872	.270	.642
Pyphen	.839	.875	.475	.591
HypheNN	.909	.910	.822	.871
BiLSTM-CRF	**.939**	**.978**	**.936**	**.984**

Table 4: Evaluation of Syllabification Systems on Wiktionary (German) and CELEX (English).

	German	EPG64	FORB	PROS
# correct lines	3431	1098	1084	1564
# faulty lines	58	114	49	173

Table 5: Size of manually annotated corpora. Faulty lines denotes the number of lines where the automatic syllabification failed. Correct lines are used for experiments, since only there the gold annotation aligns.

4.1.2 POS tagging

Since we are dealing with historical data, POS taggers trained on current data might degrade in quality and it has been frequently noted that poetry makes use of non-canonical syntactic structures (Gopidi and Alam, 2019). For German, we evaluate the robustness of POS taggers across different text genres. We use the gold annotation of the TIGER corpus (modern newspaper), and pre-tagged sentences from DTA, including annotated poetry (Lyrik), fiction (Belletristik) and news (Zeitung).[14] The STTS tagset is used. We train and test Conditional Random Fields (CRF)[15] to determine a robust POS model.[16] See Table 7 for an overview of the cross-genre evaluation. We find that training on TIGER is not robust to tag across domains, falling to around .8 F1-score when tested

---

[14]DTA was tagged with TreeTagger and manually corrected afterwards. http://www.deutschestextarchiv.de/doku/pos

[15]From the sklearn crf-suite

[16]As features, we use the word form, the preceding and following two words and POS tags, orthographic information (capitalization), character prefixes and suffixes of length 1, 2, 3 and 4.

POS	Noun	Adj.	Full V.	Adverb	Modal V.	Interj.	Pron.	Prep.	Konj.	Art.
Abr. POS Tag	NN	ADJ	VV	ADV	VM	ITJ	P	AP	KO	AR
Accent Ratio	.97	.89	.84	.75	.73	.55	.4–.015	.27	.23	.06

Table 6: Accent ratio for part-of-speech of German monosyllabic words (ratio of metrical stress).

against poetry and news from DTA. These results suggest that this is mainly due to (historical) orthography, and to a lesser extent due to local syntactic inversions.

Test	Train				
	TIGER	DTA	DTA+TIG.	Belletr.	Lyrik
Lyrik	.795	.949	.948	.947	**.953**
Belletristik	.837	**.956**	.954	.955	.955
DTA Zeitung	.793	**.934**	.933	.911	.900
TIGER	**.971**	.928	.958	.929	.913

Table 7: Evaluation of German POS taggers across genres. F1-scores.

For English, we test the Stanford core-nlp tagger.[17] The tagset follows the convention in the Penn TreeBank. This tagger is not geared towards historical poetry and consequently fails in a number of cases. We manually correct 50 random lines and determine an accuracy of 72%, where particularly the 'NN' tag is overused. This renders the English POS annotation unreliable for our experiments.

4.1.3 Additional Data and Format

The annotated corpora for English include: (1) The for-better-for-verse (FORB) collection[18] with around 1200 lines which was used by Agirreza-bal et al. (2016, 2019), and (2) the 1700 lines of poetry against which `prosodic`[19] (Anttila and Heuser, 2016; Algee-Hewitt et al., 2014) was evaluated (PROS). We merge these with our own (3) 1200 lines in 64 English poems (EPG64). The first two corpora were already annotated for metrical syllable stress. However, FORB does not contain readily available foot boundaries, and in PROS foot boundaries are occasionally set after each syllable.[20] Table 5 shows the number of lines in each

---

[17]https://nlp.stanford.edu/software/tagger.shtml

[18]https://github.com/manexagirrezabal/for_better_for_verse/tree/master/poems

[19]https://github.com/quadrismegistus/prosodic

[20]Additionally, FORB makes use of a `<seg>` tag to indicate syllable boundaries, so we do not derive the position of a syllable in a word. It also contains two competing annotations, `<met>` and `<real>`. The former is the supposedly proper metrical annotation, while the latter corresponds to a more natural rhythm (with a tendency to accept inversions and stress

of our datasets and the number of lines that were incorrectly segmented by our best syllabification systems.

Figure 4 shows an example line in the data layout that is used for the experiments, including the 'measure' that was derived with regular expressions from the meter line. 'Syll' is the position of the syllable in a word, 0 for monosyllaba, otherwise index starting at 1. We removed punctuation to properly render line measures, even through punctuation is a good signal for caesuras (see Figure 1).

```
#  tok    met ft pos  syll csr main smsr     measure      met_line
1  Look    + .  VB   0    .   1    iambic   i.penta.inv  +--+-+-+-+
2  on      - .  IN   0    .   0    iambic   i.penta.inv  +--+-+-+-+
3  my      - .  PRP$ 0    .   0    iambic   i.penta.inv  +--+-+-+-+
4  works   + :  NNS  0    :   2    iambic   i.penta.inv  +--+-+-+-+
5  ye      - .  PRP$ 0    .   0    iambic   i.penta.inv  +--+-+-+-+
6  Might   + :  NNP  1    .   1    iambic   i.penta.inv  +--+-+-+-+
7  y       - .  NNP  2    :   0    iambic   i.penta.inv  +--+-+-+-+
8  and     + :  CC   0    .   0    iambic   i.penta.inv  +--+-+-+-+
9  de      - .  VB   1    .   0    iambic   i.penta.inv  +--+-+-+-+
10 spair'  + :  VB   2    :   1    iambic   i.penta.inv  +--+-+-+-+
```

Figure 4: Tabular data format for experiments. Author of this line: Percy Blythe Shelley.

4.2 Accent Ratio of Part-of-Speech

Previous research has noted that part-of-speech annotation provides a good signal for the stress of words (Nenkova et al., 2007; Greene et al., 2010). To test this, we calculate the pos-accent ratio of monosyllabic words in our German annotation by dividing how often a particular part-of-speech appears stressed (+) in the corpus by how often this part-of-speech occurs in the corpus. We restrict this to monosyllabic words, as polysyllabic words typically have a lexical stress contour. The result is a hierarchy of stress that we report in Table 6. At the ends of the spectrum, we see that nouns are usually stressed, while articles are seldom stressed.

4.3 Learning Meter

To learn the previously annotated metrical values for each syllable, the task is framed as sequence classification. Syllable tokens are at the input and

---

clashes). We only chose `<real>` when `<met>` doesn't match the syllable count (ca. 200 cases), likely deviating from the setup in (Agirrezabal et al., 2016, 2019).

the respective `met` labels at the output. We test a nominal CRF (see section 4.1.2) and a BERT model as baselines and implement a BiLSTM-CRF[21] with pre-trained syllable embeddings. These embeddings were trained by splitting all syllables in the corpora from section 3, and training word2vec embeddings over syllables. This system uses three layers of size 100 for the BiLSTM and does the final label prediction with a linear-Chain CRF. Variable dropout of .25 was applied at both input and output. No extra character encodings were used (as these hurt both speed and accuracy).

We do a three fold cross validation with 80/10/10 splits and average the results, reporting results on the test set in Table 8. We evaluate prediction accuracy on syllables and the accuracy of whether the whole line was tagged correctly (line acc.). Line accuracy is especially important if we want to classify poetic verse measures.

| | English | | German | |
	syll. acc	line acc	syll. acc	line acc
CRF	.922	.478	.941	.553
BERT	.850	.371	.932	.498
BiLSTM-CRF	**.955**	**.831**	**.968**	**.877**
Agirrezabal (2019)	.930	.614	-	-
Antilla & Heuser (2016)	.894	.607	-	-

Table 8: Best Classifiers for Metrical Syllable Stress

Though not directly comparable (data composition differs), we include results as reported by Agirrezabal et al. (2019) for the English for-better-for-verse dataset. We also test the system 'prosodic' of Anttila and Heuser (2016) against our gold data (EPG64), resulting in .85 accuracy for syllables and .44 for lines. When only evaluating on lines that were syllabified to the correct length (their syllabifier), 27% of lines are lost, but on this subset it achieves .89 syllable and .61 line accuracy.

Learning the sequence of metrical syllable stress with BERT cannot compete our other models, possibly resulting from an improper syllable representation, as the word-piece tokenizer segments word chunks other than syllables.

We also experiment with framing the task as document (line) classification, where BERT should learn the verse label (e.g., iambic.pentameter) for a given sequence of words. On the small English dataset, BERT only achieves around .22 F1-macro and .42 F1-micro. We then tagged 20,000 lines of the large English corpus with a

BiLSTM-CRF model and trained BERT on this larger dataset, reaching .48 F1-macro and .62 F1-micro. In this setup, BERT detects frequent classes like iambic.pentameter or trochaic.tetrameter fairly well (.8), but it appears that this model mainly picks up on the length of lines and fails to learn measures other than iambus and trochee like dactyl or anapaest or irregular verse with inversions. This might limit experiments with transfer learning of verse measure knowledge.

4.4 Pairwise Joint Prosodic Task Learning

	met	feet	syllin	pos	csra	m.ac	emo
single	.964	.871	**.952**	.864	.912	.866	.328
+met	-	**.922**	.949	.856	.918	.869	.347
+feet	.961	-	.948	.853	.917	.863	.368
+syllin	.966	.900	-	.860	.919	.867	.330
+pos	.956	.879	**.953**	-	**.924**	**.879**	**.393**
+csra	.961	.886	.940	.855	-	.868	.364
+m.ac	.964	**.915**	.948	.865	.915	-	.354
+smsr	.965	.884	.942	.854	.918	.868	**.378**
+fmsr	**.968**	.899	.938	.858	**.926**	.868	**.395**
+m_line	.966	.882	.937	.853	.919	.868	**.398**
+all	.967	**.930**	.947	.790	.919	.870	.377

Table 9: Accuracy for Pairwise Joint Task Learning.

With the aim of learning the relationships between our different annotation layers, we performed experiments with a multi-task setup. We used the BiLSTM architecture from the previous experiment, where the sequence of syllable embedding vectors is at the input, and the respective sequence of labels at the output. We used the German dataset here, as the annotation is generally more reliable (e.g., POS). In this experiment we also try to learn the annotation of aesthetic emotions that was described for this dataset by Haider et al. (2020). Each line was annotated with one or two emotions from a set of nine emotions. Here, we only used the primary emotion label per line.

First, we trained a single task model for each annotation layer, then all tasks jointly (+all), and finally pair-wise combinations (+<auxiliary task>). In Table 9, we report the accuracy on syllable level for each main task with their respective auxiliary tasks.

Note that learning syllable-level POS does not benefit from any other task, not even the syllable position in the word, while several tasks like caesuras, main accents and emotions benefit from additional POS information. Predicting meter also degrades from an additional POS task, which possibly interfers with the syllable embeddings. Meter might be also more contextual than suggested in Table 6.

However, meter tagging slightly benefits from fine-grained verse measure labels. Interestingly, learning foot boundaries heavily benefits from jointly learning syllable stress. In a single task setup, foot boundaries are learned with .871 accuracy, but in combination with metrical stress, feet are learned with .922 acc. and in combination with main accents at .915. This might be expected, as foot groupings are dependent on the regularity of repeating metrical syllable stresses (though less dependent on main accents). However, our annotators only achieved Kappa agreement of .87 for feet. It is curious then, how the model overcomes this ambiguity. When learning all tasks jointly (+all), foot prediction even reaches .930, suggesting that feet are related to all other prosodic annotations.

We observe that the exchange between caesuras and main accents is negligible. However, caesuras benefit from POS (despite the absence of punctuation), syllable position (syllin) and global measures (msr), indicating that caesuras are integral to poetic rhythm and fairly dependent on syntax.

For emotions we find, despite the hard task (line instead of stanza), and only using syllable embeddings rather than proper word embeddings, that the single task setup is already better than the majority baseline. More importantly, we can see that jointly learning POS or verse measure benefits the emotion prediction (slightly the meter prediction itself: .97). This suggests that there might be a systematic relationship between meter and emotion.

5 Related Work

5.1 Annotation of Prosodic Features

Earlier work (Nenkova et al., 2007) already found strong evidence that part-of-speech tags, accent-ratio[22] and local context provide good signals for the prediction of word stress. Subsequently, models like MLP (Agirrezabal et al., 2016), CRFs and LSTMs (Estes and Hench, 2016; Agirrezabal et al., 2019) and transformer models (Talman et al., 2019) have notably improved the performance to predict the prosodic stress of words and syllables. Unfortunately, most of this work only evaluates model accuracy on syllable or word level, with the exception of Agirrezabal et al. (2019).

A digital resource with annotation of poetic meter was missing for New High German. For Middle High German, Estes and Hench (2016) annotated

---

[22]The ratio of how often a word form appears stressed vs. unstressed in a corpus

a metrical scheme for hybrid meter. Anttila et al. (2018) annotated main accents in political speeches. Agirrezabal et al. (2016, 2019) used the English *for-better-for-verse* and the dataset of Navarro et al. (2016), who annotated hendecasyllabic verse (11 syllables) in Spanish Golden Age sonnets. Algee-Hewitt et al. (2014) annotated 1700 lines of English poetry to evaluate their system.

5.2 Poetry Corpora & Generation

Several poetry corpora have been used in the NLP community. Work on English has strongly focused on iambic pentameter, e.g., of Shakespeare (Greene et al., 2010) or with broader scope (Jhamtani et al., 2017; Lau et al., 2018; Hopkins and Kiela, 2017). Other work has focused on specific genres like Spanish sonnets (Ruiz Fabo et al., 2020), limericks (Jhamtani et al., 2019), or Chinese Tang poetry (Zhang and Lapata, 2014). There are further resources with rhyme patterns (Reddy and Knight, 2011; Haider and Kuhn, 2018) or emotion annotation (Haider et al., 2020). Truly large corpora are still hard to find, besides the Gutenberg project for English and Textgrid and DTA for German.

6 Conclusion

We created large poetry corpora for English and German to support computational literary studies and annotated prosodic features in smaller corpora. Our evaluation shows that a multitude of features can be annotated through silent reading, including meter, main accents and caesuras, even though foot annotation can be challenging. Finally, we performed first experiments with a multi-task setup to find beneficial relations between certain prosodic tasks. Learning metrical annotation, including feet and caesuras, largely benefits from a global verse measure label, while foot boundaries also benefit from any joint learning with syllable stress and all features alltogether, even surpassing the human upper bound.

Acknowledgments

References

Manex Agirrezabal, Iñaki Alegria, and Mans Hulden. 2016. Machine learning for metrical analysis of English poetry. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 772–781, Osaka, Japan. The COLING 2016 Organizing Committee.

Manex Agirrezabal, Iñaki Alegria, and Mans Hulden. 2019. A comparison of feature-based and neural scansion of poetry. *RANLP 2019, arXiv preprint arXiv:1711.00938*.

Mark Algee-Hewitt, Ryan Heuser, Maria Kraxenberger, JD Porter, Jonny Sensenbaugh, and Justin Tackett. 2014. The stanford literary lab transhistorical poetry project phase ii: Metrical form. In *DH*.

Arto Anttila, Timothy Dozat, Daniel Galbraith, and Naomi Shapiro. 2018. Sentence stress in presidential speeches. *Lingbuzz Preprints*.

Arto Anttila and Ryan Heuser. 2016. Phonological and metrical variation across genres. In *Proceedings of the Annual Meetings on Phonology*, volume 3.

Derek Attridge. 2014. *The rhythms of English poetry*. Routledge.

MH Beissinger. 2012. Oral poetry. *The Princeton encyclopedia of poetry and poetics*, pages 978–81.

Klemens Bobenhausen. 2011. The metricalizer2–automated metrical markup of german poetry. *Current Trends in Metrical Analysis, Bern: Peter Lang*, pages 119–131.

Julian Brooke, Adam Hammond, and Graeme Hirst. 2015. Gutentag: an nlp-driven tool for digital humanities research in the project gutenberg corpus. In *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, pages 42–47.

Thomas Carper and Derek Attridge. 2020. *Meter and meaning: an introduction to rhythm in poetry*. Routledge.

Alex Estes and Christopher Hench. 2016. Supervised machine learning for hybrid meter. In *Proceedings of the Fifth Workshop on Computational Linguistics for Literature*, pages 1–8.

Nigel Fabb and Morris Halle. 2008. *Meter in poetry: A new theory*. Cambridge University Press.

Marjan Ghazvininejad, Xing Shi, Yejin Choi, and Kevin Knight. 2016. Generating topical poetry. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1183–1191.

Amitha Gopidi and Aniket Alam. 2019. Computational analysis of the historical changes in poetry and prose. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 14–22.

Erica Greene, Tugba Bodrumlu, and Kevin Knight. 2010. Automatic analysis of rhythmic poetry with applications to generation and translation. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 524–533.

Susanne Haaf, Alexander Geyken, and Frank Wiegand. 2014. The dta "base format": A tei subset for the compilation of a large reference corpus of printed text from multiple sources. *Journal of the Text Encoding Initiative*, (8).

Thomas Haider, Steffen Eger, Evgeny Kim, Roman Klinger, and Winfried Menninghaus. 2020. Po-emo: Conceptualization, annotation, and modeling of aesthetic emotions in german and english poetry. In *Proceedings of The 12th Language Resources and Evaluation Conference & arXiv:2003.07723*, pages 1652–1663.

Thomas Haider and Jonas Kuhn. 2018. Supervised rhyme detection with siamese recurrent networks. In *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 81–86.

Johann Christian August Heyse. 1827. Theoretisch-praktische grammatik oder lehrbuch zum reinen und richtigen sprechen, lesen und schreiben der deutschen sprache.

Jack Hopkins and Douwe Kiela. 2017. Automatically generating rhythmic verse with neural networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 168–178.

Harsh Jhamtani, Varun Gangal, Eduard Hovy, and Eric Nyberg. 2017. Shakespearizing modern language using copy-enriched sequence-to-sequence models. *arXiv preprint arXiv:1707.01161*.

Harsh Jhamtani, Sanket Vaibhav Mehta, Jaime Carbonell, and Taylor Berg-Kirkpatrick. 2019. Learning rhyming constraints using structured adversaries. *arXiv preprint arXiv:1909.06743*.

Sofoklis Kakouros, Joris Pelemans, Lyan Verwimp, Patrick Wambacq, and Okko Räsänen. 2016. Analyzing the contribution of top-down lexical and bottom-up acoustic cues in the detection of sentence prominence. *Proceedings Interspeech 2016*, 8:1074–1078.

Paul Kiparsky. 2020. Metered verse. *Annual Review of Linguistics*, 6:25–44.

Otto Knörrich. 1971. *Die deutsche Lyrik der Gegenwart*, volume 401. Kröner.

Maria Kraxenberger and Winfried Menninghaus. 2016. Mimological reveries? disconfirming the hypothesis of phono-emotional iconicity in poetry. *Frontiers in Psychology*, 7:1779.

Jey Han Lau, Trevor Cohn, Timothy Baldwin, Julian Brooke, and Adam Hammond. 2018. Deep-speare: A joint neural model of poetic language, meter and rhyme. *arXiv preprint arXiv:1807.03491*.

Katja Mellmann. 2007. Versanalyse. *Handbuch Literaturwissenschaft 2: Methoden und Theorien*, pages 81–97.

Winfried Menninghaus, Valentin Wagner, Eugen Wassiliwizky, Thomas Jacobsen, and Christine A Knoop. 2017. The emotional and aesthetic powers of parallelistic diction. *Poetics*, 63:47–59.

Yoonsook Mo, Jennifer Cole, and Eun-Kyung Lee. 2008. Naïve listeners' prominence and boundary perception. *Proc. Speech Prosody, Campinas, Brazil*, pages 735–738.

Borja Navarro, María Ribes Lafoz, and Noelia Sánchez. 2016. Metrical annotation of a large corpus of spanish sonnets: representation, scansion and evaluation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4360–4364.

Borja Navarro-Colorado. 2018. A metrical scansion system for fixed-metre spanish poetry. *Digital Scholarship in the Humanities*, 33(1):112–127.

Ani Nenkova, Jason Brenier, Anubha Kothari, Sasha Calhoun, Laura Whitton, David Beaver, and Dan Jurafsky. 2007. To memorize or to predict: Prominence labeling in conversational speech. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 9–16.

Allison Parrish. 2018. A Gutenberg Poetry Corpus https://github.com/aparrish/gutenberg-poetry-corpus.

Petr Plecháč. 2020. Relative contributions of shakespeare and fletcher in henry viii: An analysis based on most frequent words and most frequent rhythmic patterns. *Oxford Journal of Digital Humanities (DSH), arXiv preprint arXiv:1911.05652*.

Thomas Proisl and Peter Uhrig. 2016. SoMaJo: State-of-the-art tokenization for German web and social media texts. In *Proceedings of the 10th Web as Corpus Workshop (WAC-X) and the EmpiriST Shared Task*, pages 57–62, Berlin. Association for Computational Linguistics (ACL).

Sravana Reddy and Kevin Knight. 2011. Unsupervised discovery of rhyme schemes. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 77–82.

Nils Reimers and Iryna Gurevych. 2017. Reporting Score Distributions Makes a Difference: Performance Study of LSTM-networks for Sequence Tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 338–348, Copenhagen, Denmark.

Andrew Rosenberg. 2010. Autobi-a tool for automatic tobi annotation. In *Eleventh Annual Conference of the International Speech Communication Association*.

Ina Rösiger and Arndt Riester. 2015. Using prosodic annotations to improve coreference resolution of spoken text. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 83–88.

Pablo Ruiz Fabo, Helena Bermúdez Sabel, Clara Martínez Cantón, and Elena González-Blanco. 2020. The diachronic spanish sonnet corpus: Tei and linked open data encoding, data distribution, and metrical findings. *Digital Scholarship in the Humanities*.

T Steele. 2012. Verse and prose. *The Princeton encyclopedia of poetry and poetics*, pages 1507–1513.

Aarne Talman, Antti Suni, Hande Celikkanat, Sofoklis Kakouros, Jörg Tiedemann, and Martti Vainio. 2019. Predicting prosodic prominence from text with pre-trained contextualized word representations. *arXiv preprint arXiv:1908.02262*.

Rafael Valle, Kevin Shih, Ryan Prenger, and Bryan Catanzaro. 2020. Flowtron: an autoregressive flow-based generative network for text-to-speech synthesis. *arXiv preprint arXiv:2005.05957*.

Xingxing Zhang and Mirella Lapata. 2014. Chinese poetry generation with recurrent neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 670–680.